BAB 3
METODE PENELITIAN

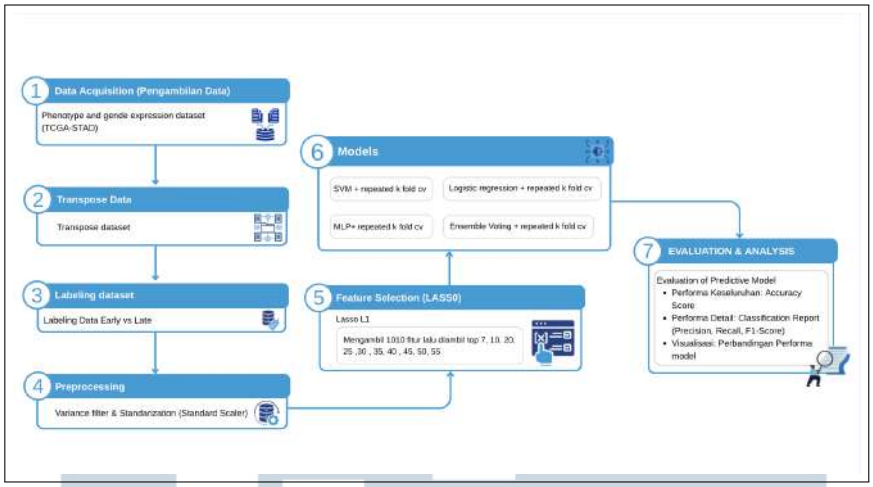
Bab ini menjelaskan metodologi penelitian yang digunakan untuk mengembangkan model klasifikasi stadium kanker lambung berbasis data ekspresi gen. Metodologi penelitian disusun secara sistematis mulai dari pengambilan data, pra-pemrosesan, seleksi fitur, pembangunan model *machine learning*, hingga evaluasi performa model. Setiap tahapan dirancang untuk memastikan hasil penelitian bersifat objektif, reprodusibel, dan relevan secara klinis.

Pendekatan yang digunakan dalam penelitian ini adalah pendekatan*machine learning* berbasis data genomik, di mana informasi biologis diekstraksi dari data ekspresi gen hasil *RNA sequencing (RNA-seq)*. Dengan memanfaatkan teknik seleksi fitur dan algoritma klasifikasi, penelitian ini bertujuan untuk menghasilkan model prediksi stadium kanker lambung yang memiliki performa optimal serta interpretabilitas biologis yang baik sebagai sistem pendukung keputusan klinis.

3.1 Alur Metodologi Penelitian

Alur metodologi penelitian ini terdiri dari beberapa tahapan utama, yaitu akuisisi data ekspresi gen, pra-pemrosesan data, seleksi fitur, pembangunan model klasifikasi, optimasi hiperparameter, serta evaluasi performa model. Alur metodologi ini dirancang untuk mengatasi karakteristik utama data ekspresi gen yang berdimensi tinggi dan memiliki rasio jumlah fitur terhadap jumlah sampel yang besar.

Gambar 3.1 menunjukkan alur metodologi penelitian secara keseluruhan.



Gambar 3.1. Alur Metodologi Penelitian

Gambar 3.1 menggambarkan alur metodologi penelitian yang diawali dengan tahap akuisisi data ekspresi gen kanker lambung. Data yang diperoleh kemudian melalui tahap pra-pemrosesan, yang meliputi pembersihan data, normalisasi, serta pemeriksaan kualitas untuk memastikan data layak digunakan dalam proses analisis.

Selanjutnya, dilakukan tahap seleksi fitur menggunakan metode regularisasi untuk mengurangi dimensi data dan mengidentifikasi gen-gen yang paling relevan terhadap status stadium kanker. Hasil seleksi fitur ini kemudian digunakan sebagai masukan pada tahap pembangunan model klasifikasi. Beberapa algoritma *machine learning* diterapkan dan dievaluasi guna mempelajari pola data serta membedakan stadium awal dan stadium lanjut kanker lambung.

Pada tahap berikutnya, optimasi hiperparameter dilakukan untuk memperoleh konfigurasi model yang optimal dengan memanfaatkan skema validasi silang. Tahap akhir dari metodologi penelitian ini adalah evaluasi performa model menggunakan metrik evaluasi yang sesuai, sehingga dapat diperoleh model klasifikasi dengan kinerja yang stabil dan dapat diandalkan.

3.2 Metode Penelitian

Penelitian ini menggunakan metode kuantitatif dengan pendekatan eksperimental. Pendekatan ini digunakan untuk mengevaluasi performa berbagai algoritma machine learning dalam mengklasifikasikan stadium kanker lambung berdasarkan data ekspresi gen. Penelitian bersifat komparatif, di mana beberapa model klasifikasi dibandingkan menggunakan metrik evaluasi yang sama untuk

menentukan model terbaik.

Pendekatan kuantitatif dipilih karena data ekspresi gen bersifat numerik dan berdimensi tinggi, sehingga memungkinkan penerapan teknik statistik dan komputasional secara optimal dalam proses analisis.

3.3 Pendekatan Metodologis

Pendekatan metodologis dalam penelitian ini terdiri dari beberapa tahapan utama, yaitu:

1. Akuisisi dan eksplorasi data ekspresi gen,
2. Pra-pemrosesan data,
3. Seleksi fitur bertahap menggunakan Lasso,
4. Pembagian data latih dan data uji,
5. Pembangunan dan optimasi model klasifikasi,
6. Evaluasi performa dan analisis hasil.

Pendekatan ini dirancang untuk mengatasi tantangan utama dalam analisis data genomik, yaitu tingginya dimensi fitur, potensi multikolinearitas antar gen, serta keterbatasan jumlah sampel.

3.4 Sumber dan Karakteristik Data

Data yang digunakan dalam penelitian ini berupa data ekspresi gen kanker lambung yang diperoleh dari basis data publik *The Cancer Genome Atlas (TCGA)*. Data ekspresi gen diperoleh melalui teknologi *RNA sequencing (RNA-seq)*, yang memungkinkan pengukuran tingkat ekspresi ribuan gen secara simultan pada setiap sampel.

Selain data molekuler, data klinis pasien juga digunakan untuk menentukan label stadium kanker berdasarkan sistem *TNM*. Untuk keperluan klasifikasi, stadium kanker dikelompokkan menjadi dua kelas, yaitu stadium awal (stadium I dan II) dan stadium lanjut (stadium III dan IV). Pengelompokan ini dilakukan untuk mendukung analisis klasifikasi biner yang relevan dalam konteks pengambilan keputusan klinis.

Berdasarkan hasil pengelompokan tersebut, jumlah sampel pada kelas stadium lanjut (kelas 1) sebanyak 249 pasien, sedangkan jumlah sampel pada kelas stadium awal (kelas 0) sebanyak 221 pasien. Distribusi ini menunjukkan bahwa data relatif seimbang antara kedua kelas, sehingga mendukung proses pelatihan dan evaluasi model klasifikasi.

3.5 Pra-pemrosesan Data

Pra-pemrosesan data dilakukan untuk memastikan kualitas dan konsistensi data sebelum digunakan dalam proses pemodelan. Tahapan ini bertujuan untuk meminimalkan potensi bias serta kesalahan yang dapat memengaruhi performa model klasifikasi. Dengan pra-pemrosesan yang tepat, data ekspresi gen dapat dianalisis secara lebih akurat dan andal.

Tahapan pra-pemrosesan yang dilakukan dalam penelitian ini meliputi penyelarasan data, penanganan nilai hilang, normalisasi dan standarisasi, serta transformasi label. Setiap tahapan dirancang untuk mengatasi karakteristik khusus data ekspresi gen yang berdimensi tinggi. Proses ini juga berperan penting dalam meningkatkan stabilitas dan generalisasi model.

3.5.1 Penyelarasan Data

Penyelarasan data dilakukan untuk memastikan bahwa setiap sampel pada data ekspresi gen memiliki pasangan data klinis yang sesuai. Proses ini bertujuan untuk menghindari ketidaksesuaian antara fitur genetik dan label stadium kanker. Sampel yang tidak memiliki informasi klinis lengkap dieliminasi untuk menjaga konsistensi dan validitas analisis.

Selain itu, penyelarasan data membantu memastikan bahwa setiap baris data merepresentasikan satu individu pasien secara utuh. Hal ini sangat penting dalam penelitian berbasis data klinis dan molekuler. Dengan demikian, kesalahan interpretasi akibat ketidaksesuaian data dapat diminimalkan.

3.5.2 Penanganan Nilai Hilang

Nilai hilang pada data ekspresi gen ditangani menggunakan metode imputasi median. Metode ini dipilih karena bersifat robust terhadap keberadaan *outlier* yang umum ditemukan pada data biologis. Selain itu, imputasi median tidak secara signifikan mengubah distribusi data asli.

Penanganan nilai hilang diperlukan untuk memastikan bahwa seluruh fitur dapat digunakan dalam proses pemodelan. Kehadiran nilai hilang dapat menyebabkan kegagalan algoritma pembelajaran mesin atau menurunkan performa model. Oleh karena itu, tahapan ini menjadi bagian penting dalam pra-pemrosesan data.

3.5.3 Normalisasi dan Standarisasi

Normalisasi dan standarisasi dilakukan untuk menyamakan skala antar fitur pada data ekspresi gen. Pada penelitian ini digunakan metode *StandardScaler* yang mengubah data agar memiliki rata-rata nol dan simpangan baku satu. Pendekatan ini membantu mengurangi dominasi fitur dengan skala besar terhadap proses pembelajaran model.

Standarisasi sangat penting untuk algoritma yang sensitif terhadap skala data, seperti *Support Vector Machine* dan *Multi-Layer Perceptron*. Tanpa standarisasi, fitur dengan nilai besar dapat mendistorsi hasil pelatihan model. Dengan demikian, proses ini berkontribusi pada peningkatan stabilitas dan akurasi klasifikasi.

3.5.4 Transformasi Label

Label stadium kanker lambung pada data klinis awalnya terdiri dari beberapa kategori stadium. Untuk mendukung pendekatan klasifikasi biner, label tersebut dikonversi menjadi dua kelas, yaitu stadium awal dan stadium lanjut. Transformasi ini dilakukan untuk menyederhanakan permasalahan klasifikasi dan meningkatkan kejelasan interpretasi hasil.

Pendekatan klasifikasi biner memungkinkan model fokus pada perbedaan utama antara dua kondisi klinis yang signifikan. Selain itu, transformasi label membantu meningkatkan stabilitas model ketika jumlah sampel pada tiap kelas tidak seimbang. Dengan demikian, hasil klasifikasi diharapkan menjadi lebih konsisten dan relevan secara klinis.

3.6 Seleksi Fitur

3.6.1 Seleksi Fitur Menggunakan LASSO

Seleksi fitur pada penelitian ini dilakukan menggunakan metode *Least Absolute Shrinkage and Selection Operator (LASSO)*. *LASSO* merupakan teknik regularisasi berbasis penalti L1 yang mampu melakukan seleksi fitur secara otomatis dengan mengecilkan koefisien fitur yang kurang relevan hingga bernilai nol. Dengan mekanisme ini, hanya gen-gen yang memiliki kontribusi signifikan terhadap proses klasifikasi yang dipertahankan.

Metode *LASSO* sangat sesuai untuk data ekspresi gen yang berdimensi tinggi dan memiliki korelasi antar fitur, karena mampu mengurangi kompleksitas model sekaligus meningkatkan interpretabilitas biologis. Dalam penelitian ini, *LASSO* diimplementasikan menggunakan model *logistic regression* dengan regularisasi L1, sehingga proses seleksi fitur dilakukan secara langsung dalam kerangka model klasifikasi.

Pada tahap awal, *LASSO* diterapkan pada seluruh fitur gen hasil preprocessing untuk mengidentifikasi gen-gen dengan koefisien non-nol. Gen-gen tersebut dianggap sebagai fitur yang lolos seleksi awal dan relevan terhadap pembedaan kelas *Early Stage* dan *Late Stage* kanker lambung.

Selanjutnya, berdasarkan nilai absolut koefisien yang dihasilkan oleh *LASSO*, dilakukan pengurutan gen dari yang memiliki kontribusi paling besar hingga yang paling kecil. Dari hasil pengurutan tersebut, dilakukan beberapa skenario pemilihan jumlah fitur, yaitu sebanyak 7, 10, 20, 25, 30, 35, 40, 45, 50, dan 55 gen teratas.

Setiap subset fitur yang dihasilkan kemudian digunakan sebagai input pada pembangunan dan evaluasi model klasifikasi stadium kanker lambung. Pendekatan ini memungkinkan analisis yang komprehensif terhadap pengaruh jumlah fitur hasil seleksi *LASSO* terhadap performa model, serta membantu menentukan jumlah fitur optimal yang mampu memberikan keseimbangan terbaik antara akurasi klasifikasi dan kompleksitas model.

3.7 Pembagian Data

Data yang telah melalui tahapan seleksi fitur selanjutnya dibagi menjadi data latih dan data uji. Pembagian data dilakukan menggunakan rasio 80% untuk data latih dan 20% untuk data uji. Data latih digunakan untuk membangun dan

mengoptimasi model, sedangkan data uji digunakan untuk mengevaluasi performa model secara independen. Untuk memastikan distribusi kelas yang seimbang pada kedua subset data, pembagian data dilakukan menggunakan metode *stratified sampling*.

3.8 Pembangunan Model Klasifikasi

Model klasifikasi yang digunakan dalam penelitian ini meliputi:

1. *Logistic Regression* yang digunakan sebagai model baseline,
2. *Support Vector Machine* (SVM) dengan kernel *Radial Basis Function* (RBF),
3. *Random Forest* sebagai model ensemble berbasis pohon keputusan,
4. *Multi-Layer Perceptron* (MLP) sebagai representasi model jaringan saraf.

Selain penggunaan model tunggal, pendekatan *ensemble learning* dengan metode *soft voting* juga diterapkan untuk mengombinasikan prediksi probabilistik dari beberapa model terbaik, dengan tujuan meningkatkan stabilitas dan performa klasifikasi secara keseluruhan.

3.9 Optimasi Hiperparameter

Optimasi hiperparameter dilakukan untuk memperoleh konfigurasi model yang menghasilkan performa terbaik. Pada penelitian ini, optimasi dilakukan menggunakan metode *Grid Search* dengan skema *k-fold cross-validation*. Proses ini bertujuan untuk mengurangi risiko *overfitting* serta meningkatkan kemampuan generalisasi model terhadap data baru.

3.10 Evaluasi Model

Evaluasi performa model dilakukan menggunakan beberapa metrik, yaitu akurasi, presisi, recall, *F1-score*, dan *Area Under the Curve (AUC)* dari kurva *Receiver Operating Characteristic (ROC)* [16]. Metrik *AUC-ROC* digunakan karena mampu memberikan gambaran performa model secara menyeluruh tanpa bergantung pada ambang batas klasifikasi tertentu.

Selain itu, *confusion matrix* digunakan untuk menganalisis jenis kesalahan klasifikasi yang dilakukan oleh model, khususnya kesalahan dalam membedakan stadium awal dan stadium lanjut kanker lambung.

3.11 Analisis Feature Importance

Analisis *feature importance* dilakukan untuk mengevaluasi kontribusi relatif setiap gen terhadap hasil prediksi model. Analisis ini bertujuan untuk meningkatkan interpretabilitas biologis serta memberikan wawasan mengenai gen yang berpotensi berperan dalam progresi kanker lambung.

3.12 Analisis Pengaruh Seleksi Fitur

Analisis ini dilakukan untuk mengevaluasi pengaruh seleksi fitur terhadap performa dan kompleksitas model. Perbandingan dilakukan antara model yang menggunakan seluruh fitur gen dan model yang menggunakan fitur hasil seleksi untuk menilai efektivitas pipeline yang diusulkan.

3.13 Pipeline Eksperimen

Pipeline eksperimen dalam penelitian ini dirancang untuk memastikan konsistensi dan reprodusibilitas hasil. Setiap algoritma klasifikasi diuji menggunakan pipeline yang sama, mulai dari standarisasi data, seleksi fitur, pelatihan model, optimasi hiperparameter, hingga evaluasi akhir pada data uji.

Seluruh eksperimen dilakukan menggunakan bahasa pemrograman Python dengan pustaka *scikit-learn*.

3.14 Ringkasan Metodologi

Metodologi penelitian ini dirancang untuk mengatasi tantangan analisis data ekspresi gen yang berdimensi tinggi dalam klasifikasi stadium kanker lambung. Dengan mengintegrasikan teknik seleksi fitur dan algoritma machine learning, penelitian ini diharapkan mampu menghasilkan model klasifikasi yang akurat, stabil, dan memiliki potensi aplikasi sebagai sistem pendukung keputusan klinis.